High-dimensional regression adjustments in randomized experiments

Stefan Wager[a,b,1], Wenfei Du[a], Jonathan Taylor[a], and Robert J. Tibshirani[a,c,1]

[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Operations, Information & Technology, Stanford Graduate School of Business, Stanford University, Stanford, CA 94305; and [c]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305

We study the problem of treatment effect estimation in randomized experiments with high-dimensional covariate information and show that essentially any risk-consistent regression adjustment can be used to obtain efficient estimates of the average treatment effect. Our results considerably extend the range of settings where high-dimensional regression adjustments are guaranteed to provide valid inference about the population average treatment effect. We then propose cross-estimation, a simple method for obtaining finite-sample–unbiased treatment effect estimates that leverages high-dimensional regression adjustments. Our method can be used when the regression model is estimated using the lasso, the elastic net, subset selection, etc. Finally, we extend our analysis to allow for adaptive specification search via cross-validation and flexible non-parametric regression adjustments with machine-learning methods such as random forests or neural networks.

high-dimensional confounders | randomized trials | regression adjustment

Randomized controlled trials are often considered the gold standard for estimating the effect of an intervention, as they allow for simple model-free inference about the average treatment effect on the sampled population. Under mild conditions, the mean observed outcome in the treated sample minus the mean observed outcome in the control sample is a consistent and unbiased estimator for the population average treatment effect.

However, the fact that model-free inference is possible in randomized controlled trials does not mean that it is always optimal: As argued by Fisher (1), if we have access to auxiliary features that are related to our outcome of interest via a linear model, then controlling for these features using ordinary least squares will reduce the variance of the estimated average treatment effect without inducing any bias. This line of research has been thoroughly explored: Under low-dimensional asymptotics where the problem specification remains fixed while the number of samples grows to infinity, it is now well established that regression adjustments are always asymptotically helpful—even in misspecified models—provided we add full treatment-by-covariate interactions to the regression design and use robust standard errors (2–10).

The characteristics of high-dimensional regression adjustments are less well understood. In a recent advance, Bloniarz et al. (11) show that regression adjustments are at least sometimes helpful in high dimensions: Given an "ultrasparsity" assumption from the high-dimensional inference literature, they establish that regression adjustments using the lasso (12, 13) are more efficient than model-free inference. This result, however, leaves a substantial gap between the low-dimensional regime—where regression adjustments are always asymptotically helpful—and the high-dimensional regime where we have only special-case results.

In this paper, we show that high-dimensional regression adjustments to randomized controlled trials work under much greater generality than previously known. We find that any regression adjustment with a free intercept yields unbiased estimates of the treatment effect. This result is agnostic as to whether the regression model was obtained using the lasso, the elastic net (14), subset selection, or any other method that satisfies this criterion.

We also propose a simple procedure for building practical confidence intervals for the average treatment effect.

Furthermore, we show that the precision of the treatment effect estimates obtained by such regression adjustments depends only on the prediction risk of the fitted regression adjustment. In particular, any risk-consistent regression adjustment can be made to yield efficient estimates of the average treatment effect in the sense of refs. 15–18. Thus, when choosing which regression adjustment to use, practitioners are justified in using standard model selection tools that aim to control prediction error, e.g., Mallow's Cp or cross-validation.

This finding presents a striking contrast to the theory of high-dimensional regression adjustments in observational studies. In a setting where treatment propensity may depend on covariates, simply fitting low-risk regression models to the treatment and control samples via cross-validation is not advised, as there exist regression adjustments that have low predictive error but yield severely biased estimates of the average treatment effect (19–22). Instead, special-case procedures are needed: For example, Belloni et al. (21) advocate a form of augmented model selection that protects against bias at the cost of worsening the predictive performance of the regression model. The tasks of fitting good high-dimensional regression adjustments to randomized vs. observational data thus present qualitatively different challenges.

The first half of this paper develops a theory of regularized regression adjustments with high-dimensional Gaussian designs. This analysis enables us to highlight the connection between the predictive accuracy of the regression adjustment and the precision of the resulting treatment effect estimate and also to considerably improve on theoretical guarantees available in prior work. In the second half of the paper, we build on these insights to develop cross-estimation, a practical method for inference about average treatment effects that can be paired with either high-dimensional regularized regression or nonparametric machine-learning methods.

Significance

As datasets get larger and more complex, there is a growing interest in using machine-learning methods to enhance scientific analysis. In many settings, considerable work is required to make standard machine-learning methods useful for specific scientific applications. We find, however, that in the case of treatment effect estimation with randomized experiments, regression adjustments via machine-learning methods designed to minimize test set error directly induce efficient estimates of the average treatment effect. Thus, machine-learning methods can be used out of the box for this task, without any special-case adjustments.

Author contributions: S.W. designed research; S.W., W.D., J.T., and R.J.T. performed research; S.W. analyzed data; and S.W. and R.J.T. wrote the paper.

Reviewers: W.L., Columbia University; and D.S., Wharton School, University of Pennsylvania.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: swager@stanford.edu or tibs@stanford.edu.

1. Setting and Notation

We frame our analysis in terms of the Neyman–Rubin potential outcomes model (23, 24). Given n i.i.d. observations (X_i, Y_i, W_i), $i = 1, 2, \ldots, n$, we posit potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$; then, the outcome that we actually observe is $Y_i = Y_i^{(W_i)}$. We focus on randomized controlled trials, where W_i is independent of all pretreatment characteristics,

$$\left\{ X_i, Y_i^{(0)}, Y_i^{(1)} \right\} \perp W_i. \qquad [1]$$

We take the predictors to be generated as $X_i \sim F(\cdot) \in \mathbb{R}^p$ and assume a homoskedastic linear model in each arm,

$$Y_i = c^{(W_i)} + X_i \cdot \beta^{(W_i)} + \varepsilon_i^{(W_i)}, c^{(w)} \in \mathbb{R}, \beta^{(w)} \in \mathbb{R}^p, \qquad [2]$$

for $w = 0, 1$, where $\varepsilon_i^{(W_i)}$ is mean-zero noise with variance σ^2; more general models are considered later. We use the notation $n_0 = |\{i : W_i = 0\}|$ and $n_1 = |\{i : W_i = 1\}|$. We study inference about the average treatment effect $\tau = \mathbb{E}[Y(1) - Y(0)]$. In our analysis, it is sometimes also convenient to study estimation of the conditional average treatment effect

$$\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \Big| X_i \right] \qquad [3]$$

$$= \bar{X} \cdot \left(\beta^{(1)} - \beta^{(0)} \right) + c^{(1)} - c^{(0)}.$$

As discussed by ref. 17, good estimators for $\bar{\tau}$ are generally good estimators for τ and vice versa. In the homogeneous treatment effects model $Y_i = c + X_i \cdot \beta + W_i \tau + \varepsilon_i$, τ and $\bar{\tau}$ coincide.

All technical derivations can be found in the *Supporting Information*, *A. Proofs*.

2. Regression Adjustments with Gaussian Designs

Suppose that we have obtained parameter estimates $\hat{c}^{(w)}$, $\hat{\beta}^{(w)}$, $w \in \{0,1\}$ for the linear model Eq. 2 via the lasso, the elastic net, or any other method. We then get a natural estimator for the average treatment effect:

$$\hat{\tau} = \bar{X} \cdot \left(\hat{\beta}^{(1)} - \hat{\beta}^{(0)} \right) + \hat{c}^{(1)} - \hat{c}^{(0)}. \qquad [4]$$

In the case where $\hat{\beta}^{(w)}$ is the ordinary least-squares estimator for $\hat{\tau}$, the behavior of this estimator has been carefully studied by refs. 8 and 10. Our goal is to characterize its behavior for generic regression adjustments $\hat{\beta}^{(w)}$, all while allowing the number of predictors p to be much larger than the sample size n.

The only assumption that we make on the estimation scheme is that it be centered: For $w \in \{0,1\}$,

$$\bar{Y}_w = \bar{X}_w \cdot \hat{\beta}^{(w)} + \hat{c}^{(w)}; \qquad [5]$$

i.e., the mean of the predicted outcomes matches that of the observed outcomes; and $\hat{\beta}^{(w)}$ is translation invariant and depends only on

$$\mathcal{F}_\beta = \left\{ X_i - \bar{X}_{W_i}, Y_i - \bar{Y}_{W_i}, W_i \right\}_{i=1}^n. \qquad [6]$$

Here, $\bar{X}_w$ and $\bar{Y}_w$ denote the mean of the outcomes Y_i and features X_i over all observations with $W_i = w$. Algorithmically, a simple way to enforce this constraint is to first center the training samples $X_i \rightarrow X_i - \bar{X}_{W_i}$, $Y_i \rightarrow Y_i - \bar{Y}_{W_i}$, run any regression method on these centered data, and then set the intercept using Eq. 5; this is done by default in standard software for regularized regression, such as glmnet (25). We also note that ordinary least-squares regression is always centered in this sense, even after common forms of model selection.

Now, if our regression adjustment has a well-calibrated intercept as in Eq. 5, then we can write Eq. 4 as

$$\hat{\tau} = \bar{X} \cdot \left(\hat{\beta}^{(1)} - \hat{\beta}^{(0)} \right) + \left(\hat{c}^{(1)} - \hat{c}^{(0)} \right)$$
$$= \bar{Y}_1 - \bar{Y}_0 + \left(\bar{X} - \bar{X}_1 \right) \cdot \hat{\beta}^{(1)} - \left(\bar{X} - \bar{X}_0 \right) \cdot \hat{\beta}^{(0)}. \qquad [7]$$

To move forward, we focus on the case where the data-generating model for (X_i, Y_i) is Gaussian; i.e., $X_i \sim \mathcal{N}(m, \Sigma)$ for some $m \in \mathbb{R}^p$ and positive-semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$, and $Y_i - \mathbb{E}[Y_i | X_i, W_i] \sim \mathcal{N}(0, \sigma^2)$. For our purpose, the key fact about Gaussian data is that the mean of independent samples is independent of the within-sample spread; i.e.,

$$\left\{ X_i - \bar{X}_{W_i}, Y_i - \bar{Y}_{W_i} \right\}_{i=1}^n \perp \left\{ \bar{X}_0, \bar{X}_1, \bar{Y}_0, \bar{Y}_1 \right\}, \qquad [8]$$

conditionally on the treatment assignments $W_1, \ldots, W_n$. Thus, because $\hat{\beta}^{(w)}$ depends only on the centered data $X_i - \bar{X}_{W_i}$ and $Y_i - \bar{Y}_{W_i}$, we can derive a simple expression for the distribution of $\hat{\tau}$. The following is an exact finite sample result and holds no matter how large p is relative to n; a key observation is that $\bar{X} - \bar{X}_w$ is mean zero by randomization of the treatment assignment, for $w = 0, 1$.

Proposition 1. *Suppose that our regression scheme for $\hat{c}^{(w)}$ and $\hat{\beta}^{(w)}$ is centered and that our data-generating model is Gaussian as above. Then, writing $\|v\|_\Sigma^2 := v^\top \Sigma v$ for $v \in \mathbb{R}^p$,*

$$\hat{\tau} - \bar{\tau} | n_0, n_1, \hat{\beta}^{(0)}, \hat{\beta}^{(1)} \overset{d}{=} \mathcal{N}(0, A),$$

$$A = \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \left(\sigma^2 + \left\| \hat{\bar{\beta}} - \bar{\beta} \right\|_\Sigma^2 \right), \qquad [9]$$

$$\bar{\beta} = \frac{n_1 \beta^{(0)} + n_0 \beta^{(1)}}{n}, \quad \hat{\bar{\beta}} = \frac{n_1 \hat{\beta}^{(0)} + n_0 \hat{\beta}^{(1)}}{n}.$$

If the errors in $\hat{\beta}^{(0)}$ and $\hat{\beta}^{(1)}$ are roughly orthogonal, then

$$\left\| \hat{\bar{\beta}} - \bar{\beta} \right\|_\Sigma^2 \approx \frac{n_1^2}{n^2} \left\| \hat{\beta}^{(0)} - \beta^{(0)} \right\|_\Sigma^2 + \frac{n_0^2}{n^2} \left\| \hat{\beta}^{(1)} - \beta^{(1)} \right\|_\Sigma^2 \qquad [10]$$

and, in any case, twice the right-hand side is always an upper bound for the left-hand side. Thus, the distribution of $\hat{\tau}$ effectively depends on the regression adjustments $\hat{\beta}^{(w)}$ only through the excess predictive error

$$\left\| \hat{\beta}^{(w)} - \beta^{(w)} \right\|_\Sigma^2 = \mathbb{E}\left[\left((X - m) \cdot \left(\hat{\beta}^{(w)} - \beta^{(w)} \right) \right)^2 \Big| \hat{\beta}^{(w)} \right], \qquad [11]$$

where the above expectation is taken over a test set example X. This implies that, in the setting of *Proposition 1*, the main practical concern in choosing which regression adjustment to use is to ensure that $\hat{\beta}^{(w)}$ has low predictive error.

The above result is conceptually related to recent work by Berk et al. (3) (also ref. 26), who showed that the accuracy of low-dimensional covariate adjustments using ordinary least-squares regression depends on the mean-squared error of the regression fit; they also advocate using this connection to provide simple asymptotic inference about τ. Here, we showed that a similar result holds for any regression adjustment on Gaussian designs, even in high dimensions; and in the second half of this paper we discuss how to move beyond the Gaussian case.

Risk Consistency and the Lasso. As stated, *Proposition 1* provides the distribution of $\hat{\tau}$ conditionally on $\hat{\beta}^{(w)}$ and so is not directly comparable to related results in the literature. However, whenever $\hat{\beta}^{(w)}$ is risk consistent in the sense that

$$R\left(\hat{\beta}^{(w)} \right) := \left\| \hat{\beta}^{(w)} - \beta^{(w)} \right\|_\Sigma^2 \rightarrow_p 0, \qquad [12]$$

for $w = 0, 1$, we can asymptotically omit the conditioning.

Theorem 2. *Suppose that, under the conditions of Proposition 1, we have a sequence of problems where $\hat{\beta}^{(w)}$ is risk consistent (Eq. 12), and $\mathbb{P}[W=1] \to \pi$. Then,*

$$\sqrt{n}(\hat{\tau} - \overline{\tau}) \Rightarrow \mathcal{N}\left(0, \frac{\sigma^2}{\pi(1-\pi)}\right), \qquad [13]$$

or, in other words, $\hat{\tau}$ is efficient for estimating $\overline{\tau}$ (15–18).

In the case of the lasso, *Theorem 2* lets us substantially improve over the best existing guarantees in the literature (11). The lasso estimates $\hat{\beta}^{(w)}$ as the minimizer over β of

$$\sum_{\{i:W_i=w\}} \frac{1}{2}\left(Y_i - \bar{Y}_w - (X_i - \bar{X}_w) \cdot \beta\right)^2 + n_w\lambda\|\beta\|_1, \qquad [14]$$

for some penalty parameter $\lambda > 0$. Typically, the lasso is used when we believe a sparse regression adjustment to be appropriate. In our setting, it is well known that the lasso satisfies $R(\hat{\beta}^{(w)}) = \mathcal{O}_P\left(\|\Sigma\|_{\mathrm{op}}^2 \|\beta^{(w)}\|_0 \log(p)/n_w\right)$, provided the penalty parameter λ is well chosen and Σ does not allow for too much correlation between features (27, 28).

Thus, whenever we have a sequence of problems as in *Theorem 2* where $\beta^{(w)}$ is k-sparse, i.e., $\beta^{(w)}$ has at most k nonzero entries, and $k\log(p)/n \to 0$, we find that $\hat{\tau}$ is efficient in the sense of Eq. 13. Note that this result is much stronger than the related result of ref. 11, which shows that lasso regression adjustments yield efficient estimators $\hat{\tau}$ in an ultrasparse regime with $k \ll \sqrt{n}/\log(p)$.

To illustrate the difference between these two results, it is well known that if $k \ll \sqrt{n}/\log(p)$, then it is possible to do efficient inference about the coefficients of the underlying parameter vector β (29–31), and so the result of ref. 11 is roughly in line with the rest of the literature on high-dimensional inference. Conversely, if we have only $k \ll n/\log(p)$, accurate inference about the coefficients of β is in general impossible without further conditions on the covariance of X (32, 33). However, we have shown that we can still carry out efficient inference about τ. In other words, the special structure present in randomized trials means that much more is possible than in the generic high-dimensional regression setting.

Inconsistent Regression Adjustments. Even if our regression adjustment $\hat{\beta}^{(w)}$ is not risk consistent, we can still use *Proposition 1* to derive unconditional results about $\hat{\tau}$ whenever

$$R(\widehat{\bar{\beta}}) := \left\|\widehat{\bar{\beta}} - \bar{\beta}\right\|_{\Sigma}^2 \to_p R_\infty. \qquad [15]$$

We illustrate this phenomenon in the case of ridge regression, where regression adjustments generally reduce—but do not eliminate—excess test-set risk. Recall that ridge regression estimates $\hat{\beta}^{(w)}$ as the minimizer over β of

$$\sum_{\{i:W_i=w\}} \frac{1}{2}\left(Y_i - \bar{Y}_w - (X_i - \bar{X}_w) \cdot \beta\right)^2 + n_w\lambda\|\beta\|_2^2. \qquad [16]$$

The following result relies on random-matrix theoretic tools for analyzing the predictive risk of ridge regression (34).

Theorem 3. *Suppose we have a sequence of problems in the setting of Proposition 1 with $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$, such that the spectrum of the covariance Σ has a weak limit. Following ref. 34, suppose moreover that the true parameters $\beta^{(0)}$ and $\beta^{(1)}$ are independently and randomly drawn from a random-effects model with*

$$\mathbb{E}[\beta^{(w)}] = 0 \quad and \quad \mathrm{Var}[\beta^{(w)}] = \frac{\alpha^2}{p}I_{p\times p}, \quad with \ \alpha > 0. \qquad [17]$$

Then, selecting $\hat{\beta}^{(w)}$ in Eq. 4 via ridge regression tuned to minimize prediction error, and with $\mathbb{P}[W=1] \to \pi$, we get $\sqrt{n}(\hat{\tau} - \overline{\tau}) \Rightarrow \mathcal{N}(0, S)$,

$$S = 2\sigma^2 + \frac{\alpha^2}{\gamma}\left(\frac{\pi}{v_0\left(-\frac{\gamma\sigma^2}{\alpha^2(1-\pi)}\right)} + \frac{1-\pi}{v_1\left(-\frac{\gamma\sigma^2}{\alpha^2\pi}\right)}\right), \qquad [18]$$

where the $v_w(-\lambda)$ are the companion Stieltjes transforms of the limiting empirical spectral distributions for the treated and control samples, as defined in the proof.

To interpret this result, we note that the quantity $v_w(-\lambda)$ can also be induced via the limit (35, 36)

$$\frac{1}{n_w}\mathrm{tr}\left(\left(\frac{1}{n_w}\sum_{\{i:W_i=w\}} X_iX_i^\top + \lambda I_{n_w\times n_w}\right)^{-1}\right) \to_p v_w(-\lambda), \quad \text{for } \lambda > 0.$$

Finally, we note that the limiting variance of $\hat{\tau} - \overline{\tau}$ obtained via ridge regression above is strictly smaller than the corresponding variance of the unadjusted estimator, which converges to $(\sigma^2 + (\pi^2 + (1-\pi^2))\alpha^2\mathrm{tr}(\Sigma)/p)/(\pi(1-\pi))$; this is because optimally tuned ridge regression strictly improves over the "null" model $\hat{\beta}^{(w)} = 0$ in terms of its predictive accuracy.

3. Practical Inference with Cross-Estimation

In the previous section, we found that—given Gaussianity assumptions—generic regression adjustments yield unbiased estimates of the average treatment effect and also that low-risk regression adjustments lead to high-precision estimators. Here, we seek to build on this insight and to develop simple inferential procedures about τ and $\overline{\tau}$ that attain the above efficiency guarantees, all while remaining robust to deviations from Gaussianity or homoskedasticity.

Our approach is built around cross-estimation, a procedure inspired by data splitting and the work of refs. 37 and 38. We first split our data into K equally sized folds (e.g., $K=5$ or 10) and then, for each fold $k = 1, \ldots, K$, we compute

$$\hat{\tau}^{(k)} = \bar{Y}_1^{(k)} - \bar{Y}_0^{(k)} + \left(\bar{X}^{(k)} - \bar{X}_1^{(k)}\right) \cdot \hat{\beta}^{(1, -k)} \\ - \left(\bar{X}^{(k)} - \bar{X}_0^{(k)}\right) \cdot \hat{\beta}^{(0, -k)}. \qquad [19]$$

Here, $\bar{Y}_1^{(k)}$, $\bar{Y}_0^{(k)}$, etc., are moments taken over the kth fold, whereas $\hat{\beta}^{(1, -k)}$ and $\hat{\beta}^{(0, -k)}$ are centered regression estimators computed over the $K-1$ other folds. We then obtain an aggregate estimate $\hat{\tau} = \sum_{k=1}^K \hat{\tau}^{(k)} n^{(k)}/n$, where $n^{(k)}$ is the number of observations in the kth fold. An advantage of this construction is that an analog to the relation Eq. 8 now automatically holds, and thus our treatment effect estimator $\hat{\tau}$ is unbiased without assumptions. Note that the result below references both the average treatment effect τ and the conditional average treatment effect $\overline{\tau}$.

Theorem 4. *Suppose that we have n independent and identically distributed samples satisfying Eq. 1, drawn from a linear model Eq. 2, where X_i has finite first moments and the conditional variance of $Y_i^{(w)}$ given X_i may vary. Then, $\mathbb{E}[\hat{\tau}|X_1, \ldots, X_n] = \overline{\tau}$. If, moreover, the $\hat{\beta}^{(w, -k)}$ are all risk consistent in the sense of Eq. 12 for $k = 1, \ldots, K$, and both the signals $X_i \cdot \beta^{(w)}$ and residuals $Y_i - \mathbb{E}[Y_i|X_i, W_i = w]$ are asymptotically Gaussian when averaged, then writing $\overline{\sigma}_w^2 = \mathbb{E}[\mathrm{Var}[Y_i^{(w)}|X_i]]$, we have*

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}\left(0, \frac{\overline{\sigma}_0^2}{1-\pi} + \frac{\overline{\sigma}_1^2}{\pi} + \left\|\beta^{(1)} - \beta^{(0)}\right\|_\Sigma^2\right). \qquad [20]$$

In the homoskedastic case, i.e., when the variance of $Y^{(w)}$ conditionally on X does not depend on X, then the above is efficient. With heteroskedasticity, the above is no longer efficient because we are in a linear setting and so inverse-variance weighting could improve precision; however, Eq. 20 can still be used as the basis for valid inference about τ.

Confidence Intervals via Cross-Estimation. Another advantage of cross-estimation is that it allows for moment-based variance estimates for $\hat{\tau}$. Here, we discuss practical methods for building confidence intervals that cover the average treatment effect τ. We can verify that the variance of $\hat{\tau}^{(k)}$ is V_k after conditioning on the $\hat{\beta}^{(w, -k)}$ and $n_w^{(k)}$, with

$$V_k = \sum_{w \in \{0,1\}} \frac{1}{n_w^{(k)}} \text{Var}\left[Y^{(w)} - X \cdot \widehat{\overline{\beta}}^{(-k)} \bigg| \widehat{\overline{\beta}}^{(-k)} \right]. \qquad [21]$$

Now, the above moments correspond to observable quantities on the kth data fold, so we immediately obtain a moment-based plug-in estimator $\hat{V}_k$ for V_k. Finally, we build α-level confidence intervals for τ as

$$\tau \in \hat{\tau} \pm z_{1-\alpha/2} \hat{V}, \quad \hat{V} = \sum_{k=1}^{K} \left(\frac{n^{(k)}}{n} \right)^2 \hat{V}_k, \qquad [22]$$

where $z_{1-\alpha/2}$ is the appropriate standard Gaussian quantile. In the setting of *Theorem 4*, i.e., with risk consistency and bounded second moments, we can verify that the $\hat{\tau}^{(k)}$ are asymptotically uncorrelated and so the above confidence intervals are asymptotically exact.

Cross-Validated Cross-Estimation. High-dimensional regression adjustments usually rely on a tuning parameter that controls the amount of regularization, e.g., the parameter λ for the lasso and ridge regression. Although theory provides some guidance on how to select λ, practitioners often prefer to use computationally intensive methods such as cross-validation.

Now, our procedure in principle already allows for cross-validation: If we estimate $\hat{\beta}^{(0, -k)}$ in Eq. **19** via any cross-validated regression adjustment that relies only on all but the kth data folds, then $\hat{\tau}^{(k)}$ will be unbiased for τ. However, this requires running the full cross-validated algorithm K times, which can be very expensive computationally.

Here, we show how to obtain good estimates of $\hat{\tau}$ using only a single round of cross-validation. First, we specify K regression folds, and for each $k \in \{1, \ldots, K\}$ and $w \in \{0,1\}$ we compute $\bar{X}_{k,w}$ and $\bar{Y}_{k,w}$ as the mean of all observations in the kth fold with $W_i = w$. Next, we center the data such that $\tilde{X}_i = X_i - \bar{X}_{k, W_i}$ and $\tilde{Y}_i = Y_i - \bar{Y}_{k, W_i}$ for all observations in the kth fold. Finally, we estimate $\hat{\beta}^{(w, -k)}$ by running a standard out-of-the-box cross-validated algorithm (e.g., cv. glmnet for R) on the $(\tilde{X}_i, \tilde{Y}_i, W_i)$ triples with the same K folds as specified before and then use Eq. **19** to compute $\hat{\tau}$.

The actual estimator that we use to estimate $\hat{\beta}^{(0)}$ and $\hat{\beta}^{(1)}$ in our experiments is inspired by the procedure of Imai and Ratkovic (39). Our goal is to let the lasso learn shared "main effects" for the
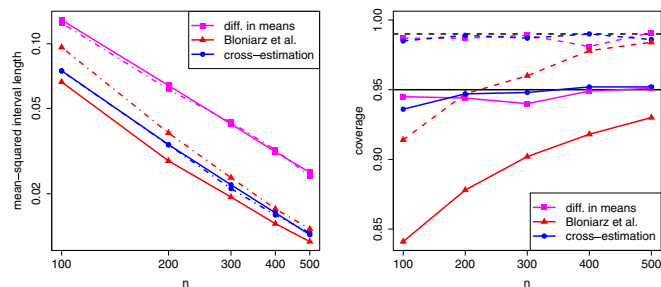
Fig. 1. Simulation results with $\beta = (1, 0, 0, \ldots, 0)$, $\mathbb{P}[W=1]=0.2$, $\rho=0$, and $p=500$. All numbers are based on 1,000 simulation replications. *Left* panel shows both the average variance estimate $\hat{V}$ produced by each estimator (solid lines) and the actual variance $\text{Var}[\hat{\tau}]$ of the estimator (dashed-dotted lines); note that $\hat{V}$ is directly proportional to the squared length of the confidence interval. *Right* panel depicts realized coverage for both 95% confidence intervals (solid lines) and 99% confidence intervals (dashed lines).
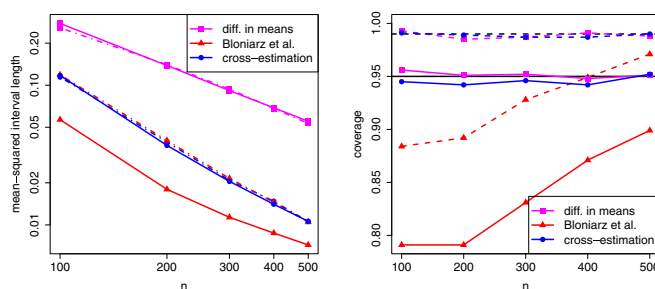
Fig. 2. Simulation results with β proportional to a permutation of $(1, 2^{-1}, 3^{-1}, \ldots, p^{-1})$, $\|\beta\|_2 = 2$, $\mathbb{P}[W=1]=0.5$, $\rho=0.8$, and $p=500$. All numbers are based on 1,000 simulation replications. The plots are produced the same way as in Fig. 1.

treatment and control groups. To accomplish this, we first run a $2p$-dimensional lasso problem,

$$\hat{\beta}, \hat{\gamma} = \text{argmin}_{\beta, \gamma} \bigg\{ \lambda(\|\beta\|_1 + \|\gamma\|_1) + \sum \left(\tilde{Y}_i - \left(\tilde{X}_i \cdot \beta + (2W_i - 1)\tilde{X}_i \cdot \gamma \right) \right)^2 \bigg\}, \qquad [23]$$

and then set $\hat{\beta}^{(0)} = \hat{\beta} - \hat{\gamma}$ and $\hat{\beta}^{(1)} = \hat{\beta} + \hat{\gamma}$. We simultaneously tune λ and estimate τ by cross-validated cross-estimation as discussed above. When all our data are Gaussian, this procedure is exactly unbiased by the same argument as used in *Proposition 1*; and even when X is not Gaussian, it appears to work well in our experiments.

4. Nonparametric Machine-Learning Methods

In our discussion so far, we have focused on treatment effect estimation using high-dimensional, linear regression adjustments and showed how to provide unbiased inference about τ under general conditions. Here, we show how to extend our results about cross-estimation to general nonparametric regression adjustments obtained using, e.g., neural networks or random forests (40). We assume a setting where

$$\mathbb{E}[Y(w) | X = x] = \mu^{(w)}(x)$$

for some unknown regression functions $\mu^{(w)}(x)$, and our goal is to leverage estimates $\hat{\mu}^{(w)}(x)$ obtained using any machine-learning method to improve the precision of $\hat{\tau}$, as[*,†]

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}^{(1, -i)}(X_i) - \hat{\mu}^{(0, -i)}(X_i) \right) + \sum_{\{i:W_i=1\}} \frac{Y_i - \hat{\mu}^{(1, -i)}(X_i)}{n_1}$$

$$- \sum_{\{i:W_i=0\}} \frac{Y_i - \hat{\mu}^{(0, -i)}(X_i)}{n_0}, \qquad [24]$$

where $\hat{\mu}^{(w, -i)}$ is any estimator that does not depend on the ith training example; for random forests, we set $\hat{\mu}^{(w, -i)}(X_i)$ to be the "out-of-bag" prediction at X_i. To motivate Eq. **24**, we start from Eq. **7** and expand out terms using the relation

---

*We note that Eq. **24** depends on $\hat{\mu}^{(0, -i)}(X_i)$ and $\hat{\mu}^{(1, -i)}(X_i)$ implicitly only through $\bar{\mu}^{(-i)}(X_i) = n_1/n\hat{\mu}^{(0, -i)}(X_i) + n_0/n\hat{\mu}^{(1, -i)}(X_i)$. It may thus also be interesting to estimate $\bar{\mu}^{(-i)}(X_i)$ directly using, e.g., the "tyranny of the minority" scheme of Lin (8).

†A related estimator is studied by Rothe in the context of classical nonparametric regression adjustments, e.g., local regression, for observational studies with known treatment propensities. [Rothe C (2016) The value of knowing the propensity score for estimating average treatment effects, (IZA), discussion paper 9989.]

$$\left(\bar{X} - \bar{X}_1\right) \cdot \hat{\beta}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}^{(1)}(X_i) - \frac{1}{n_1} \sum_{\{i:W_i=1\}} \hat{\mu}^{(1)}(X_i), \qquad [25]$$

where $\hat{\mu}^{(1)}(x) = x \cdot \hat{\beta}^{(1)} + \hat{c}^{(1)}$, etc. The remaining differences between Eq. **24** and Eq. **7** are due to the use of out-of-bag estimation to preserve randomization of the treatment assignment W_i conditionally on the corresponding regression adjustment. We estimate the variance of $\hat{\tau}$ using the formula

$$\hat{V} = \sum_{w \in \{0,1\}} \sum_{\{i:W_i=w\}} \frac{\left(Y_i - \frac{n_0}{n}\hat{\mu}^{(1,-i)}(X_i) - \frac{n_1}{n}\hat{\mu}^{(0,-i)}(X_i)\right)^2}{n_w(n_w - 1)}. \qquad [26]$$

Formal Results. The following result characterizes the behavior of this estimator, under the assumption that the estimator is "jackknife compatible," meaning that the expected jackknife estimate of variance for $\hat{\mu}^{(w)}$ converges to 0. We define this condition in *Supporting Information, A. Proofs* and verify that it holds for random forests.

Theorem 5. *Suppose that $\hat{\mu}$ is jackknife compatible. Then, the estimator $\hat{\tau}$ in Eq. **24** is asymptotically unbiased, $\mathbb{E}[\hat{\tau}|X_1, \ldots, X_n] = \bar{\tau} + o(1/\sqrt{n})$. Moreover, if the regression adjustments $\hat{\mu}_w$ are risk consistent in the sense that[‡] $1/n \sum_{i=1}^{n} (\hat{\mu}^{(w,-i)}(X_i) - \mu^{(w)}(X_i))^2 \to_p 0$, and the potential outcomes $Y_i^{(w)}$ have finite second moments, then $\hat{\tau}$ is efficient and $(\hat{\tau} - \tau)/(\hat{V})^{1/2}$ is asymptotically standard Gaussian.*

We note that there has been considerable recent interest in using machine-learning methods to estimate heterogeneous treatment effects (42–45). In relation to this literature, our present goal is more modest: We simply seek to use machine learning to reduce the variance of treatment effect estimates in randomized experiments. This is why we obtain more general results than the papers on treatment heterogeneity.

5. Experiments

In our experiments, we focus on two specific variants of treatment effect estimation via cross-estimation. For high-dimensional linear estimation, we use the lasso-based method Eq. **23** tuned by cross-validated cross-estimation. For nonparametric estimation, we use Eq. **24** with random forest adjustments. We implement our method as an open-source R package, crossEstimation, built on top of glmnet (25) and randomForest (46) for R. *Supporting Information, B. Additional Simulation Results* and Table S1 have additional simulation results.

Simulations. We begin by validating our method in a simple simulation setting with $Y = X\beta + W\tau + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$. In all simulations, we set the features X to be Gaussian with autoregressive AR-ρ covariance. We compare our lasso-based cross-estimation with both the simple difference-in-means estimate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$ and the proposal of Bloniarz et al. (11) that uses lasso regression adjustments tuned by cross-validation. Our method differs from that of Bloniarz et al. (11) in that we use a different algorithm for confidence intervals and also that we use the joint lasso algorithm Eq. **23** instead of computing separate lassos in both treatment arms.

Figs. 1 and 2 display results for different choices of β, ρ, etc., while varying n. In both cases, we see that the confidence intervals produced by our cross-estimation algorithm and the method of Bloniarz et al. (11) are substantially shorter than those produced by the difference-in-means estimator. Moreover,

[‡]With random forests, ref. 41 provides such a risk-consistency result.
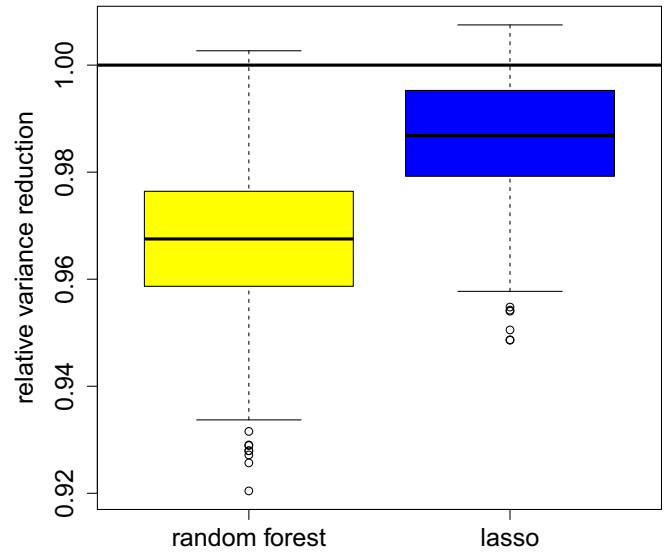
Fig. 3. Reduction in squared confidence interval length achieved by random forests and a lasso-based method, relative to the simple difference-in-means estimator. Confidence intervals rely on cross-estimation. The plot was generated using 1,000 simulation replications.

our confidence intervals accurately represent the variance of our estimator (compare solid and dashed-dotted lines in Figs. 1 and 2, *Left*) and achieve nominal coverage at both the 95% and 99% levels. Conversely, especially in small samples, the method of Bloniarz et al. (11) underestimates the variance of the method and does not achieve target coverage.

Understanding Attitudes Toward Welfare. We also consider an experimental dataset collected as a part of the General Social Survey.[§] The dataset is large ($N = 28{,}646$ after preprocessing), so we know the true treatment effect essentially without error: The fraction of respondents who say we spend too much on assistance to the poor is smaller than the fraction of respondents who say we spend too much on welfare by 0.35. To test our method, we repeatedly drew subsamples of size $n = 2{,}000$ from the full dataset and examined the ability of both lasso-based and random-forest–based cross-estimation to recover the correct answer. We had $p = 12$ regressors.

First, we note that both variants of cross-estimation achieved excellent coverage. Given a nominal coverage rate of 95%, the simple difference-in-means estimator, lasso-based cross-estimation, and random-forest cross-estimation had realized coverage rates of 96.3%, 96.5%, and 95.3%, respectively, over 1,000 replications. Meanwhile, given a nominal target of 99%, the realized numbers became 99.0%, 99.0%, and 99.3%. We note that this dataset has non-Gaussian features and exhibits considerable treatment effect heterogeneity.

Second, Fig. 3 depicts the reduction in squared confidence interval length for individual realizations of each method. More formally, we show boxplots of $\hat{V}_{\text{lasso/rf}}/\hat{V}_{\text{simple}}$, where $\hat{V}$ is the variance estimate used to build confidence intervals. Here, we see that although cross-estimation may not improve the precision of the simple method by a large amount, it consistently improves performance by a small amount. Moreover, in this example,

STATISTICS

random forests result in a larger improvement in precision than lasso-based cross-estimation.

6. Discussion

In many applications of machine-learning methods to causal inference, there is a concern that the risk of specification search, i.e., trying out many candidate methods and choosing the one that gives us a significant result, may reduce the credibility of empirical findings. This has led to considerable interest in methodologies that allow for complex model-fitting strategies that do not compromise statistical inference.

One prominent example is the design-based paradigm to causal inference in observational studies, whereby we first seek to build an "observational design" by looking only at the features X_i and the treatment assignments W_i and reveal the outcomes Y_i only once the observational design has been set (48, 49). The observational design may rely on matching, inverse-propensity

weighting, or other techniques. As the observational design is fixed before the outcomes Y_i are revealed, practitioners can devote considerable time and creativity to fine-tuning the design without compromising their analysis.

From this perspective, we have shown that regression adjustments to high-dimensional randomized controlled trials exhibit a similar opportunity for safe specification search. Concretely, imagine that once we have collected data from a randomized experiment, we provide our analyst only with class-wise centered data: W_i, $X_i - \bar{X}_{W_i}$, and $Y_i - \bar{Y}_{W_i}$. The analyst can then use these data to obtain any regression adjustment they want, which we will then plug into Eq. 4. Our results guarantee that—at least with a random Gaussian design—the resulting treatment effect estimates will be unbiased regardless of the specification search the analyst may have done using only the class-wise centered data. Cross-estimation enables us to mimic this phenomenon with non-Gaussian data.

1. Fisher R (1925) *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, UK).
2. Athey S, Imbens G (2016) The econometrics of randomized experiments. *arXiv*: 1607.00698.
3. Berk R, et al. (2013) Covariance adjustments for the analysis of randomized field experiments. *Eval Rev* 37(3–4):170–196.
4. Ding P, Feller A, Miratrix L (2016) Decomposing treatment effect variation. *arXiv*: 1605.06566.
5. Freedman D (2008) On regression adjustments in experiments with several treatments. *Ann Appl Stat* 2(1):176–196.
6. Freedman DA (2008) On regression adjustments to experimental data. *Adv Appl Math* 40(2):180–193.
7. Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47(1):5–86.
8. Lin W (2013) Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann Appl Stat* 7(1):295–318.
9. Rosenbaum PR (2002) Covariance adjustment in randomized experiments and observational studies. *Stat Sci* 17(3):286–327.
10. Cochran W (1977) *Sampling Techniques* (Wiley, New York), 3rd Ed.
11. Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B (2016) Lasso adjustments of treatment effect estimates in randomized experiments. *Proc Natl Acad Sci USA* 113(27): 7383–7390.
12. Chen S, Donoho D, Saunders M (1998) Atomic decomposition for basis pursuit. *SIAM J Sci Comput* 20(1):33–61.
13. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58: 267–288.
14. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B* 67(2):301–320.
15. Bickel P, Klaassen C, Ritov Y, Wellner J (1998) *Efficient and Adaptive Estimation for Semiparametric Models* (Springer, New York).
16. Hahn J (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2):315–331.
17. Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat* 86(1):4–29.
18. Robins JM, Rotnitzky A (1995) Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc* 90(429):122–129.
19. Athey S, Imbens GW, Wager S (2016) Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv*:1604.07125.
20. Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2016) Program evaluation with high-dimensional data. *Econometrica*, in press.
21. Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud* 81(2):608–650.
22. Farrell MH (2015) Robust inference on average treatment effects with possibly more covariates than observations. *J Econom* 189(1):1–23.
23. Neyman J (1990) On the application of probability theory to agricultural experiments. Essay on principles, section 9. translation of original 1923 paper, which appeared in *Roczniki Nauk Rolniczych. Stat Sci* 5(4):465–472.
24. Rubin DB (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *J Educ Psychol* 66(5):688–701.
25. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22.
26. Pitkin E, et al. (2013) Improved precision in estimating average treatment effects. *arXiv*:1311.0291.
27. Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann Stat* 37(4):1705–1732.
28. Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Ann Stat* 37(1):246–270.
29. Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J Mach Learn Res* 15(1):2869–2909.
30. Van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat* 42(3):1166–1202.
31. Zhang CH, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc B* 76(1):217–242.
32. Cai TT, Guo Z (2015) Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *arXiv*:1506.05539.
33. Javanmard A, Montanari A (2015) De-biasing the lasso: Optimal sample size for Gaussian designs. *arXiv*:1508.02757.
34. Dobriban E, Wager S (2015) High-dimensional asymptotics of prediction: Ridge regression and classification. *arXiv*:1507.03003.
35. Bai Z, Silverstein JW (2010) *Spectral Analysis of Large Dimensional Random Matrices* (Springer, New York), Vol 20.
36. Marchenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Math USSR Sbornik* 114(4):507–536.
37. Aronow PM, Middleton JA (2013) A class of unbiased estimators of the average treatment effect in randomized experiments. *J Causal Inference* 1(1):135–154.
38. Tibshirani R, Efron B (2002) Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol* 1(1):1–15.
39. Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* 7(1):443–470.
40. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.
41. Scornet E, Biau G, Vert JP (2015) Consistency of random forests. *Ann Stat* 43(4): 1716–1741.
42. Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci USA* 113(27):7353–7360.
43. Green DP, Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin Q* 76(3):491–511.
44. Hill JL (2012) Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 20(1):217–240.
45. Wager S, Athey S (2015) Estimation and inference of heterogeneous treatment effects using random forests. *arXiv*:1510.04342.
46. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2(3): 18–22.
47. Wager S (2016) Causal inference with random forests. PhD thesis (Stanford University, Stanford, CA).
48. Rosenbaum PR (2002) *Observational Studies* (Springer, New York).
49. Rubin DB (2007) The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med* 26(1):20–36.